August 10, 2019

Kimra Castleton

Robert Fondaw

Neeta Shrivastava

Project 2, Group 6

**Transit Systems of the World**

Extract

On Kaggle.com we located various csv files containing data on transit systems throughout the world. Upon examining and discussing the data contained within, we narrowed to two csv files with information on the cities and stations of transit systems. We then imported the csv files with Python Pandas in Jupyter Notebook.

Transform

We converted the csv files to dataframes and excluded certain nonapplicable columns. We then renamed remaining columns to better describe the content. Final columns for the cities dataframe were city\_id, city\_name, start\_year, and country. Final columns for the stations dataframe were station\_id, station\_name, buildstart, opening, and city\_id. The two dataframes shared the column City\_id.

Cleaning the data included dropping null values from the stations dataframe. City dataframe had no nulls. When we counted the number of dropped rows it accounted for ten percent of the station data so the decision was made to keep the nulls and have them accounted for after the ETL process.

Load

We created a database name transit\_systems in PGAdmin to use PostgreSQL and added to the script to create the city and station tables. In cities table we established City\_ID as the primary key and Station\_ID as the primary key in the stations table.

A master table was created by merging the cities and stations tables to make it easier for the analyst performing calculations with the data after the ETL process. Station\_ID was kept as the primary key in the master joined table.

We experimented with basic queries such as select all and count to assure database usability. Finally we grouped by, then ordered by country, and pulled the station count, to maximize and demonstrate the data’s usefulness.